Computers and Society Research Journal (2025), **1**
https://doi.org/10.54822/QVHM1662

ARTICLE

The Utrecht Housing dataset: A housing appraisal dataset

Sieuwert van Otterloo and Pavlo Burda

Email: sieuwert@ictinstitute.nl

(First published online February 1, 2025)

Abstract

This paper introduces a real-world dataset for analysing and predicting house prices. The dataset consists of actual data on the Dutch housing market collected in 2024 for a total of 153 houses in one city (Utrecht in The Netherlands). The dataset incorporates diverse variables on individual houses, including property characteristics (e.g., house type, build year, geolocation, area, energy label) and market metrics (e.g., asking price, final price). The data was collected from two public sources. The dataset has been created to help researchers and educators to demonstrate machine learning principles on several problem types. It can be used for classification (energy label and energy efficiency) and regression/ price estimation. There are ten original input features and one derived feature. The dataset can be freely used without restrictions under a Creative Commons license and is available via open data platform Kaggle.

Keywords: dataset, housing, real estate, Netherlands, Appraisal, Regression, Classification

Introduction

Data science is an interdisciplinary field that combines statistical techniques, computational methods, and domain expertise. This discipline has fundamentally changed how individuals and organizations gather information and drive their decisions. However, the availability of sufficiently high quality data has always been a limiting factor to the use of data science.

Digital technology is being used in almost all aspects of our society and this has led to a huge increase in the amount of available data. This offers many opportunities for doing data science. At the same time, the incorrect or inappropriate use of data can also pose risk to society. To capture these opportunities and help people identify and mitigate risks, it is important for all graduates to learn the principles of data science.

To teach data science, educators need to have access to a variety of datasets and example applications. In our educational activities, we noticed an absence of small, high quality datasets with up-to-date, realistic data and applications. The existing smaller datasets were not suitable to show students how data science can be applied, and large datasets often require significant preprocessing efforts before they can be used successfully. We therefore created a new dataset that any educator can use as a realistic example of data science application.

Our dataset is called the Utrecht housing dataset, version 2.0. It replaces version 1.0, which was a synthetic dataset with generated data. The new version 2.0 consists of real data collected from multiple public sources. The dataset can be used to demonstrate how data science can be used

by small teams or even individuals on a real-world problem. The size of the dataset is small (153 datapoints) but sufficient to apply several data science techniques. The dataset has been carefully reviewed to remove any data quality issues, allowing users to focus on applications instead of working on data quality. The dataset has been created around a common real-world problem that will be familiar to many people: how to determine the value of an item you wish to buy.

The dataset is released as creative commons, and can be used freely for any purpose. If you use it, please cite this paper. The dataset is available on Kaggle: https://www.kaggle.com/datasets/ictinstitute/utrecht-housing-dataset.

Housing dataset background and related work

The housing market is an interesting market to study, since finding a suitable house is a relatable concern for everyone. Many people have experienced living in different houses and will experience the problem of deciding what to bid on an available house. Deciding to buy or even a house is one of the biggest financial decisions people make and many people are therefore willing to collect data. In many markets, the government aims to make the market more or less fair by enforcing transparency and fairness through regulation. The regulation for instance covers what information must be given to buyers, how ownership is transferred and offering public records of previous transactions. Data on individual houses also has additional use cases, for example, the data can be used for city planning purposes, where one can try to use machine learning to predict house features such as energy efficiency labels [HvB23].

Predicting house prices is an example of a more general class of problems: predicting the value of individual items. This is an interesting case for data scientists in business, because it is challenging and has practical value.

Related datasets used in education

There are several excellent introductory text books on data science and machine learning, such as, those by Geron [Gér22] and Vanderplas [Van16]. Many of these books use smaller datasets to illustrate visualisation and analysis methods.

Historically, a small number of datasets such as the Iris dataset [UK21] have been used in many books. The Iris dataset has even been called the “Hello World” example of data science and machine learning [UK21]. Its continuing use shows the need for smaller datasets to explain fundamental methods. The small size of the dataset encourages users to actually inspect and review outliers, something that is harder to do for massive datasets but that is important when using data science in practice. This is a useful first dataset but it is not related to any practical business problem. In our view, one needs additional datasets to fully explain data science and its applications.

The Titanic survivors dataset is another well known dataset that, formalized by Dawson [Daw95], has been used in statistics before 1995. It shows how one can use a variety of features to make predictions, in this case, predicting the probability of survival. It is useful for explaining basic machine learning principles, but it is not related to any practical use case: such predictions cannot be used in any business setting.

Another well known dataset worth mentioning is the MNIST database consisting of sample handwritten letters [LBBH98]. It was intended for benchmarking various machine learning [BHN19]. It is an excellent dataset for computer vision education and research. It is however only useful for computer vision and not for price prediction or regression.

Finally, there are multiple housing-related datasets that are used by multiple authors. A good well-known housing dataset is the California housing dataset [Gér22]. This dataset is for instance used in the second chapter of Aurélien Géron’s book ‘Hands-On Machine learning with Scikit-Learn

and TensorFlow' [Gér22]. This is an interesting and useful dataset to experiment with deep learning models [Che24]. Similarly, the AMES dataset provide interesting housing features [DC11][1]. One limitation of the California housing dataset is that it contains median house prices: it provides macro-economic insights but cannot be used to show how to predict prices of individual items. Since making predictions for individual items is an important application of data science and machine learning, the Utrecht Housing dataset was designed to contain details on individual houses.

An older but not good dataset is the Boston Housing dataset. The Boston Housing dataset was compiled by David Harrison Jr. Daniel L. Rubinfeld for their 1975 paper on Hedonistic pricing [HR78]. It became a standard for regression analysis and feature engineering research [CCWB21, BG19] and was even included in modern libraries such as scikit-learn and tensorflow. The dataset contains two problematic columns: 'B' for impact of black population and 'LSTAT' for percentage of lower status people in the area [Fai22]. These columns aim to clarify some hard to measure environmental issues, but are based on the presence of black people. This is inaccurate , and it also seems racially motivated or based on historical racial ideas that are no longer acceptable. Either way, this is not a good dataset for introducing people to data science or statistics. It should not be used and any book or paper using this dataset without any warnings should not be used either. The dataset was included in standard data science libraries such as scikit-learn but it has been deprecated. One should use other datasets such as the California Housing dataset or the Utrecht Housing dataset.

Collection details

The dataset was created with the following requirements. First, we wanted to create a dataset based on data that is commonly used by people when making house purchase decisions. Second, we wanted to make a small dataset (100-200 datapoints) to avoid running into performance and memory issues, but large enough to investigate relationships between variables and create detailed visualizations. We also decided to select houses from a relatively small area (one city and a few neighbouring towns) since many people are indeed often searching and comparing houses in one city. While the majority of houses in the dataset belong to one of Utrecht's districts, we decided to include a few houses just outside the Utrecht municipality. This allows researchers and students to measure any effects on house value of being officially in Utrecht.

An important design decision was whether to make an effort to collect actual purchase prices, since such information is harder to obtain than list prices. Eventually we decided to include real prices from actual purchases, and not use asking price as a proxy for value. Just relying on asking prices is not a good idea when making an important purchase decision as we will explain later in this paper. Based on these requirements and familiarity with Dutch data sources, we decided to collect data for the the city of Utrecht. Utrecht is large enough that there are no immediate privacy risks. We decided to use the Dutch land registry (Kadaster)[2] as a data source for the transaction prices, i.e., the final house price. Upon a small fee, the land registry provides an extract in PDF format of all purchases in the last 20-ish years for a specific zipcode. The Kadaster document reports the transaction price of a property, last valuation date and the lot area (if any). No other data is provided, one needs additional sources to obtain features of interest. To collect further data we decided to use a public listing website for houses. There are multiple websites for houses in The Netherlands, such as Funda, Huislijn and Pararius. We decided to use the best known website, Funda [3]. On Funda one can search for available houses but also for previous listings that were already sold. We used this functionality to search for individual zipcodes with multiple sold houses. Given a selected zipcode, all the web pages of each sold house were saved in the PDF format.

[1] Ames Housing Dataset - https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset
[2] Kadaster - https://www.kadaster.nl/winkel
[3] Funda - https://www.funda.nl/

We then extracted the relevant features from each saved PDF, and collated them with the data from Kadaster, specifically, the variables `lot-area`, `retailvalue` and `valuationdate`. In case of incomplete data, i.e. when the data was not available from Kadaster yet, or the additional data was no longer available on Funda, the house was not included in the dataset.

Finally we enriched the dataset by adding geographic location data using a third datasource. The coordinates of each house (`x-coor` and `y-coor`) were derived from their full address listed on the PDFs by means of an online address-to-GPS service[4]. Similarly, the `dist-from-train` variable has been derived by computing the direct flight distance in kilometers (Euclidean distance). The data was then reviewed for any potential quality issues. We removed a few houses with a size larger than 250 square meters, since they would be visible as outliers in most visualisations. The resulting geographical distribution is shown in Figure 1: the dataset covers multiple locations around Utrecht and its outskirts.

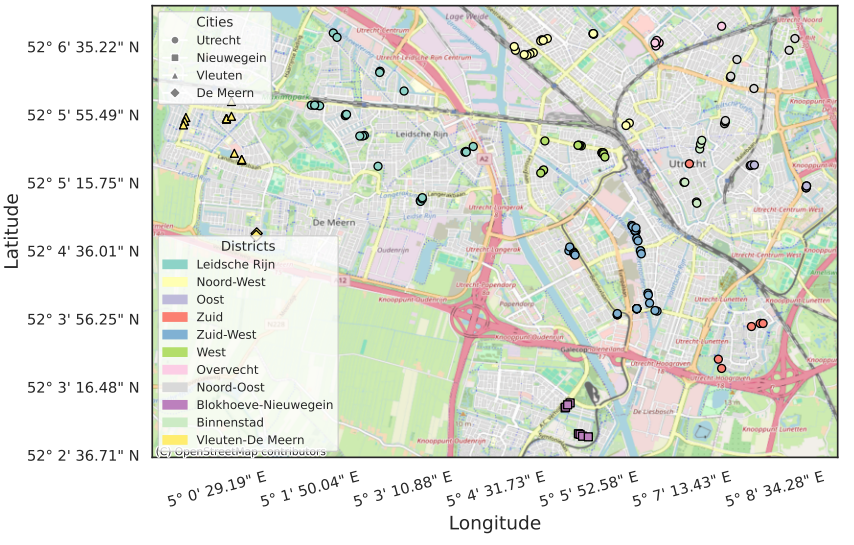


Figure 1. Scatterplot of X and Y coordinates on the map of Utrecht, showing approximate locations. The points are shape-coded by city and color-coded by district. It is clearly visible that the 153 houses in the dataset cover most of Utrecht and Vleuten, while Nieuwegein and De Meern are only partially covered.

Dataset overview

The dataset is a sample of houses sold in Utrecht in or around 2024. The data was collected from two sources: the digital marketplace and the Dutch land registry. The observations are geographically distributed across 9 of the 10 districts (wijken) inside Utrecht city borders, as seen in Table 1.

The dataset is provided as a CSV file. Each line contains data for one house. The values are separated by commas. The complete explanation of each column follows in Table 2. The order of elements is: `id`, `zipcode4`, `zipcode6`, `zipcode6id`, `housetype`, `lot-area`, `house-area`, `garden-size`, `rooms`, `bathrooms`, `x-coor`,`y-coor`, `buildyear`, `retailvalue`, `askingprice`, `energylabel`, `energyeff`, `valuationdate`, `street`, `subdistrict` `district`, `city`, `dist-from-train`. There are in total 153 observations, of which 10 have an unspecified garden area (‘garden-size’) and 3 unspecified energy label (‘energylabel’ and ‘energy-eff’) because they were not available.

[4]GPS coordinaten - https://www.gps-coordinaten.nl/

Table 1. Distribution of observations over districts and inhabitants

District	Obs. (#,%)		Inhabitants (#,%)	
Binnenstad	7	4.52%	19581	5,38%
Nieuwegein (city)	8	5.16%	65974	not Utrecht
Leidsche Rijn	25	16.13%	42783	11,0%
Noord-Oost	9	5.81%	39510	11,1%
Noord-West	16	10.32%	44925	12,6%
Oost	7	4.52%	32203	9,22%
Overvecht	5	3.23%	34152	9,72%
Vleuten-De Meern	17	10.97%	50502	13,8%
West	15	9.68%	29258	8,34%
Zuid	7	4.52%	27769	7,87%
Zuid-West	37	23.87%	38672	10,9%

Privacy and FAIR data considerations

By publishing this dataset, we want to promote the distribution and use of data according to the FAIR Data principles [WDA+16]. FAIR is an acronym for Findable, Accessible, Interoperable, Reusable. The FAIR Data principles aim to make sure data is reusable by other researchers and thus contribute to making scientific research easier to reproduce, extend or apply. The Utrecht housing dataset is findable on the open platform Kaggle and via this publication. The data is freely available without restrictions or restrictive licensing conditions. To make the data interoperable and reusable, we tried to use clearly defined values for each feature and omitted hard to interpret features.

One challenge when collecting or publishing data, is respecting the right to privacy of individuals that the data relates to. In Europe, the collection of data related to natural persons is restricted by the GDPR [Cou16]. Data collection for research purposes by Dutch universities is further restricted by the Netherlands Code of Conduct for research integrity [KNN+18]. This code of conduct requires researchers to obtain explicit permission from the data subject to process their personal data. We avoided the collection of non-public personal data in this project, in order to comply with these principles. We did not publish any of the available interior photos, nor any of the lengthy textual descriptions that might contain personal data. We also took care to select streets where multiple houses have been sold, so that the street names by itself do not refer to individual houses. Finally, we replaced the actual house numbers with 001, 002 etc. We decided to leave the actual street names and locations in. The street names function as a convenient name during discussion of outliers, and the locations are needed for visualisations and for linking to other datasets.

To make sure that the validity of each datapoint can be checked, we saved all source documents as PDF. The actual listing pages and the actual property records are available for review upon request. These will however not be published or distributed since they may contain personal data. The data collection was performed with respect to the fair use principle, whereby the web requests to the services were performed manually, few at a time, and modulated over several weeks to avoid any service impediments.

Data Exploration Results

This section provides example results for the exploratory data analysis, starting with Table 3: the table shows the descriptive statistics for the numerical variables in the dataset. For example, the average number of rooms in a house is 4, within a reasonable standard deviation of 1.6. On the

Table 2. Variables description.

Variable	Description	Count	Example
id	Unique id between 5000 and 9000.	153	5034
zipcode4	Dutch 4-digit postal code. Roughly corresponds to a sub-district.	27	3522
zipcode6	Full Dutch 6-character postal code (4 digits, 2 letters). Corresponds to part of a street.	72	3522EG
zipcode6id	Same as zipcode6 with trailing last three digits of `id`	153	3522EG034
housetype	Type of house, either 'woonhuis' (residential house) or 'appartement' (flat).	2	woonhuis
lot-area	Total area of the land plot the house is built on in square meters. 'appartement' type of houses have 0 lot-area.	62	113
house-area	Living area of the house in square meters. 30 square meters is a tiny house, 200 square meters is a mansion.	90	137
garden-size	The size of the garden in square meters.	57	50
rooms	Number of rooms.	9	7
bathroom	Number of bathrooms. Most houses have one bathroom. Some houses have 2 bathrooms.	2	1
x-coor	Latitude in decimal degrees of the house position. It is rounded to 4 decimals to retain precision at the level of individual street or large building.	110	52.0669
y-coor	Longitude in decimal degrees of the house position. It is rounded to 4 decimals.	119	5.1144
buildyear	The construction year. All houses but one were built in 20th and 21st century. The oldest house is from 1320.	57	1933
retailvalue	The actual transaction value as registered in the Dutch land registry in thousands of euros.	123	787
askingprice	The price for which the house was offered in 2024	79	650
energylabel	The EU energy efficiency label ranges from G to A++. A home with energy label A (or higher) is energy efficient, while a home with energy label G is not energy efficient.	10	A
energyeff	A binary number (1 or 0) for very energy efficient house, i.e., energy label is A or higher	2	1
valuationdate	The official transaction date when the final transaction value is registered in land registry. Follows ISO 8601 format[a].	98	2024-10-02
street	Street name.	47	Socrateslaan
subdistrict	One of the 33 Utrecht sub-districts and neighboring towns.	26	Dichterswijk-Rivierenwijk
district	One of the 10 major Utrecht districts including neighboring towns.	11	Zuid-West
city	The township (gemeente). Some houses are located in sub-urban townships close to Utrecht.	4	Utrecht
dist-from-train	Crow's flight distance from Utrecht central train station in kilometers.	113	2.44

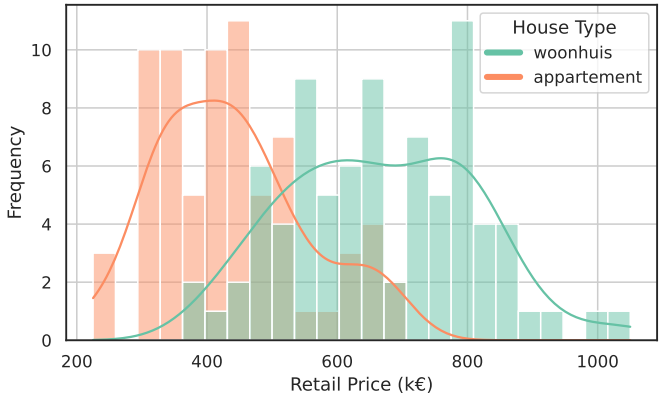
[a]ISO 80601 - https://en.wikipedia.org/wiki/ISO_8601

other hand, the average lot area is $71m^2$ but it varies greatly as the standard deviation is more than its mean ($80m^2$), likely due to apartments having a 0 lot area.

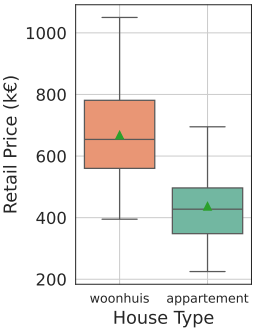
Figure 2 shows the distribution of the house prices over the house type (`woonhuis` and `appartement`). In the histogram of Figure 2a, we see that the retail price distribution of residential houses is slightly skewed to the right with respect to the apartments. Indeed, as shown the box plots of

Table 3. Descriptive statistics of numerical variables

	lot-area	house-area	rooms	bathrooms	retailvalue	askingprice	dist-from-train
mean	69.50	98.83	4.03	1.09	559.30	506.14	3.05
std	79.21	35.67	1.52	0.29	173.82	162.08	1.88
min	0.00	28.00	1.00	1.00	225.00	225.00	0.68
25%	0.00	76.00	3.00	1.00	425.00	375.00	1.63
50%	48.00	93.00	4.00	1.00	541.00	490.00	2.68
75%	125.00	118.00	5.00	1.00	676.00	625.00	4.32
max	400.00	204.00	8.00	2.00	1050.00	995.00	7.57



(a) Histogram with smoothing function that represents the distribution of retail value of residential homes and apartments (color-coded). The distributions have different centers and both approximate the shape of the normal distribution.



(b) Boxplots representing the distributions of retail value by house type. This representation allows to visualize the price means around which the house prices concentrate.

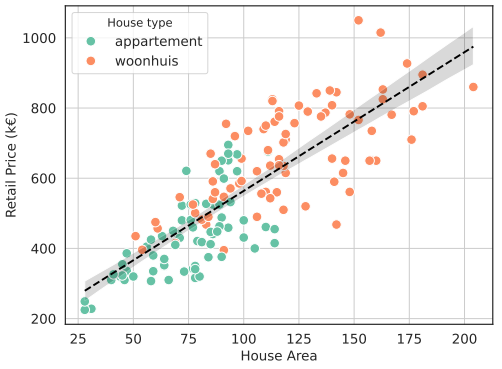
Figure 2. Distribution of retail value by house type.

Figure 2b, we can observe the difference of the price means between the two: a typical residential house costs on average around €600k, while an apartment is around €400k. To know whether this mean difference is significant for Utrecht houses and apartments, we can test if the mean difference is not due to chance with the Welsh's t-test [McN21]. We can go on with this statistic as the two shapes in Figure 2a still resemble a normal distribution. The Welsh's t-test yields: 11.1 and p-value: 2.7^{-17}. The p-value of well below 0.05 indicates that there is a significant price difference between the house types. Therefore, we can say that the precise mean difference is €231k and it ranges between €190k and €272k (its confidence interval).

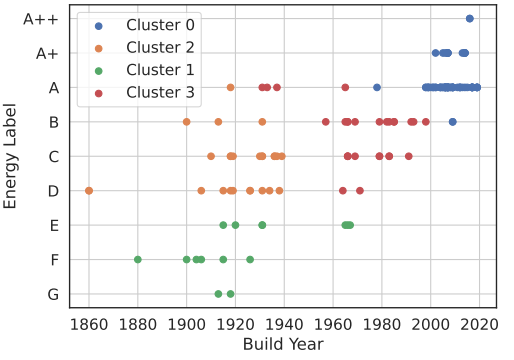
Figure 3a visualizes the correlation between house area and the retail value by scattering the datapoints over the two variables. By computing the correlation between the two variables (Pearson's correlation) we obtain: $r = 0.811$. This signals a high correlation between the two variables, therefore we can try fitting a trend line in Figure 3a (by means of linear regression). The linear relationship indicates a reasonable trend of an increasing house price to follow an increased house area.

Figure 3b compares the build year and the energy labels. We used these two features to cluster houses into 4 groups using K-means by following the elbow method to select the number of clusters [Tho53]. The clusters we obtained can be described as:

- cluster 0: modern, energy-efficient homes
- cluster 1: least efficient homes, mostly before 1940s



(a) Scatterplot of house area and retail value color-coded by type of property. A linear regression approximates the positive relation between the variables with error bands (in gray).



(b) Visualizing the correlation between build year and energy label. The K-means clustering into 4 groups highlights houses after 2000 forming a cohesive cluster 0 with efficient houses, with houses pre-2000s' less efficient. Clusters 1 and 3 can be considered as pre-1940 not renovated and renovated housing. An outlier with year 1320 was removed for readability.

Figure 3. Visualizing correlation examples.

- cluster 2: improved housing from before the 1940s
- cluster 3: aging, mid-level efficient housing

Recall Figure 1 which provides a scatter plot of house locations (X-Y coordinates) on Utrecht's map. It is easy understand the geographical distribution and the covered areas of Utrecht and, thanks to the colors, we can distinguish also the various districts too. The marker's shape also helps to visualize the towns of Vleuten-De Meren and Nieuwegein with shaped dots occurring more often at the west and south peripheries. Similarly, Figure 4 shows houses on Utrecht's map, but this time the color gradient represents the build year whereby we can notice that the farther we go from the city center (binnestad), the newer are the houses.

Possible Use Cases

As our goal includes learning about data science, some fitting research tasks are outlined below.

1) Exploratory Data Analysis. This step allows to practice exploratory visualizations, such as histograms, box plots, scatter plots, correlation heatmaps, and draw preliminary insights to guide further analysis.

- Identify the distribution of house sizes, locations and prices.
- Visualize price trends across different neighborhoods or districts.
- Examine correlations between features (e.g., house size, number of bedrooms, house age) and final selling price.
- Determine relations between house area and location, house area and build year, or rooms and bathrooms.
- Analyse the relation between retail value and asking price to explain how common overbidding was.

2) Prediction Models. In this step, the goal is to build a model to predict the features that are relevant according to the domain expert. A very typical prediction problem would be to try to predict retail price or asking price. These values must be chosen when listing a house for sale or making a bid. Note that when you are predicting retail price, you probably do not want to use asking price as an input and vice versa since these are related and often unknown when starting to list a house. It

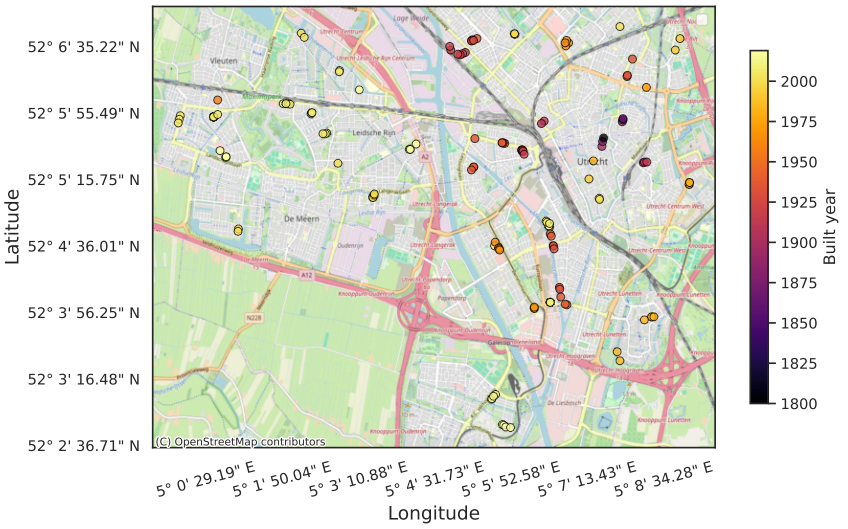


Figure 4. Scatterplot of X-Y coordinates on Utrecht's map showing houses locations and build year. The points coded with a color gradient from older to newer houses. The progression of newer houses is evident by looking from the center to the outskirts of the city. An outlier with year 1320 was removed for readability reasons.

is possible to achieve this by using linear regression [CS13] or more advanced methods, with due assumptions. One can formulate many interesting variations on this main prediction problem, a few of these are listed below.

- Try making two separate models, one for `appartement` and one for `woonhuis`. Compare the performance of this approach against a combined model.
- Try making two separate models, one for energy-efficient houses and one for non-efficient houses. Compare the performance of this approach against a combined model.
- Use x-coor and y-coor to create new features, such as distance to schools, train station, parks, coffee factory etc. Try to find a correlation to house value for each distance.
- Try to make predictions using only zipcode4 and house-area. Then rank the zipcodes based on the effect of each zipcode on the predicted value. Try to see if the ranking changes when you add build year or garden size.
- Separate the houses based on the decade that they were built in and determine which decade has the highest average retail value. Check whether you can explain the differences in average value based on the number of rooms, house-area or lot-area.
- Evaluate the application of tree-based methods (e.g., Random Forest, Gradient Boosted Trees) and understand feature importance and non-linear relationships [BRAN19].
- Instead of directly predicting retail value, compute the value per square meter of house area and try to predict this value. Note that you need to decide how to apply your performance metrics.

Another goal is to classify the energy efficiency of houses, specifically the energy label as an example of multi class classification, and energy efficiency (`energyeff`) as an example of binary classification. Suitable techniques would be logistic regression and decision trees.

One can also build a decision tree to determine the city based on x-coor and y-coor. A small tree should work. For a bigger challenge, build a decision tree for zipcode4 based on x- and y-coor.

3) Asking price vs. Final Price. An interesting feature of this dataset are the asking price and retail price variables. These signal that there may be differences in the advertised house price and what the buyers are actually willing to pay. This offers an insight into the current (as for 2024) real estate market trends, such as offer-demand dynamics.

- Formulate a custom target variable (i.e., the negotiation gap).
- Build a model to predict the negotiation gap based on property features (e.g., what variables influence the most the negotiation gap?).
- Identify which property groups/locations have the smallest/largest negotiation gap.

4) Clustering Techniques. Another feature of this dataset concerns house characteristics whereby houses can be grouped together by various features. For example, an un-intuitive yet interesting combination can be the distance from train station and garden size. For it, cluster individual properties to identify areas or house groups with similar characteristics.

- Choose relevant house attributes.
- Choose relevant clustering algorithms (e.g., K-Means, Hierarchical clustering, DBSCAN) [Van16].
- Evaluate cluster quality and characterize each cluster (e.g., "affordable housing", "luxury district", what clustering parameters were used?, etc.).

5) Hypothesis Testing. Imagine to perform a research task where you need to prove beyond reasonable doubt a relation between two variables. You can do so with hypothesis testing [McN21].

- Think about one or two hypotheses, (e.g., "Homes with more than one bathroom have a higher average selling price" or "The average negotiation gap is significantly different in the city center vs. more distant districts".
- Test the hypothesis by conducting t-tests, ANOVA, or non-parametric tests to tell whether the observed differences are *not* due to randomness.
- Compute confidence intervals, e.g., for mean prices for homes with and without gardens, to measure the uncertainty of your estimates with this dataset.

7) Recommendation System for House Buyers. As an example of applying dataset-specific knowledge and data science techniques to a real-word use case, one can create a simplified house recommendation system based on buyer preferences (e.g., location, house type, budget).

- The recommendation system can be implemented as a web application.
- You can integrate additional domain specific factors, such as distance to other points of interest (schools, parks, supermarkets, etc).
- You can implement a collaborative filtering whereby recommendations can be influenced by other users' previous choices, such as popular user choices with similar preferences (requires simulated users or the app to be publicly accessible).

Conclusions

The paper presents a real-world dataset from the 2024 Dutch housing market, offering a valuable resource for data science applications. Compiled manually with an emphasis on accuracy, the dataset supports tasks such as regression, classification, and clustering. Its design adheres to FAIR principles, ensuring accessibility and ethical use. With features enabling analysis of individual property characteristics and pricing dynamics, the dataset is well-suited for educational and research purposes, providing an alternative to synthetic or outdated datasets.

Acknowledgement

The authors would like to thank all colleagues from the Lectoraat AI at the Utrecht University of Applied Sciences and the teaching assistants and other participants of the Utrecht Summerschool in Data Science and Machine Learning, held in 2022, 2023 and 2024. The participation and feedback from all the summerschool were essential in identifying what datasets are required for data science education.

References

[BG19] Brad Boehmke and Brandon M Greenwell. *Hands-on machine learning with R*. Chapman and Hall/CRC, 2019. https://doi.org/10.1201/9780367816377.

[BHN19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[BRAN19] Mousumi Banerjee, Evan Reynolds, Hedvig B Andersson, and Brahmajee K Nallamothu. Tree-based analysis: a practical approach to create clinical decision-making tools. *Circulation: Cardiovascular Quality and Outcomes*, 12(5):e004879, 2019.

[CCWB21] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. A historical perspective of explainable artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, 11(1), 2021.

[Che24] Audrey Chen. Deep Learning in Real Estate Prediction: An Empirical Study on California House Prices. *The National High School Journal of Science*, 2024.

[Cou16] Council of European Union. General data protection regulation, 2016. http://data.europa.eu/eli/reg/2016/679/oj.

[CS13] Samprit Chatterjee and Jeffrey S Simonoff. *Handbook of regression analysis*. John Wiley & Sons, 2013.

[Daw95] Robert J MacG Dawson. The “unusual episode” data revisited. *Journal of Statistics Education*, 3(3), 1995.

[DC11] Dean De Cock. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

[Fai22] Fairlearn. Revisiting the Boston Housing Dataset. https://fairlearn.org/main/user_guide/datasets/boston_housing_data.html, 2022. Accessed: 2025-01-07.

[Gér22] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O’Reilly Media, Inc., 2022.

[HR78] David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

[HvB23] Sanne Hettinga, Rein van ’t Veer, and Jaap Boter. Large scale energy labelling with models: The eu tabula model versus machine learning with open data. *Energy*, 264:126175, 2023.

[KNN+18] KNAW, NFU, NWO, TO2-federatie, Vereniging Hogescholen, and VSNU. Nederlandse gedragscode wetenschappelijke integriteit, 2018.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[McN21] Keith McNulty. *Handbook of regression modeling in people analytics: With examples in R and Python*. Chapman and Hall/CRC, 2021. https://peopleanalytics-regression-book.org/.

[Tho53] Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[UK21] Antony Unwin and Kim Kleinman. The iris data set: In search of the source of virginica. *Significance*, 18(6):26–29, 2021. https://academic.oup.com/jrssig/article/18/6/26/7038520.

[Van16] Jake VanderPlas. *Python data science handbook: Essential tools for working with data*. " O’Reilly Media, Inc.", 2016. https://jakevdp.github.io/PythonDataScienceHandbook/.

[WDA+16] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.